Benchmarking the Raw Model-Generated Background Forecast in Rapidly Updated Surface Temperature Analyses. Part II: Gridded Benchmark

THOMAS M. HAMILL

Physical Sciences Division, NOAA/Earth System Research Laboratory, Boulder, Colorado

MICHAEL SCHEUERER

Cooperative Institute for Research in the Environmental Sciences, University of Colorado Boulder, Boulder, Colorado

(Manuscript received 7 February 2019, in final form 11 November 2019)

ABSTRACT

This is the second part of a series on benchmarking raw 1-h high-resolution numerical weather prediction surface-temperature forecasts from NOAA's High-Resolution Rapid Refresh (HRRR) system. Such 1-h forecasts are commonly used to underpin the background for an hourly updated surface temperature analysis. The benchmark in this article was produced through a gridded statistical interpolation procedure using only surface observations and a diurnally, seasonally dependent gridded surface temperature climatology. The temporally varying climatologies were produced by synthesizing high-resolution monthly gridded climatologies of daily maximum and minimum temperatures over the contiguous United States with yearly and diurnally dependent estimates of the station-based climatologies of surface temperature. To produce a 1-h benchmark forecast, for a given hour of the day, say 0000 UTC, the gridded climatology was interpolated to station locations and then subtracted from the observations. These station anomalies were statistically interpolated to produce the 0000 UTC gridded anomaly. This anomaly pattern was continued for 1 h and added to the 0100 UTC gridded climatology to generate the 0100 UTC gridded benchmark forecast. The benchmark is thus a simple 1-h persistence of the analyzed deviations from the diurnally dependent climatology. Using a cross-validation procedure with July 2015 and August 2018 data, the gridded benchmark provided competitive, relatively unbiased 1-h surface temperature forecasts relative to the HRRR. Benchmark forecasts were lower in error and bias in 2015, but the HRRR system was highly competitive or better than the gridded benchmark in 2018. Implications of the benchmarking results are discussed, as well as potential applications of the simple benchmarking procedure to data assimilation.

1. Introduction

Very short-term numerical weather predictions (NWP) of surface (2 m) temperature may exhibit systematic biases and increased errors due to the very substantial challenges of modeling the state of the land surface and its interactions with the atmosphere. These deficiencies have several potential causes, including land surface temperature and soil moisture initial state error and bias (Holmes et al. 2012; de Rosnay et al. 2014), misestimated predictions of the downward solar radiation reaching the surface (Yang et al. 2006; Paquin-Ricard et al. 2010; Räisänen and Järvinen 2010; Thelen and Edwards 2013; Van Weverberg et al. 2015; Ruiz-Arias et al. 2016), misestimated predictions of mechanical mixing of surface air with air further aloft (Sandu et al. 2013), and mismodeling of the land–atmosphere feedbacks (Dirmeyer et al. 2018). Some of the bias in the validation against observations, however, may be due to the comparison of model-forecast data intended to represent box averages against observations with characteristics unique to each particular site (i.e., representativeness error).

High-resolution, hourly surface temperature analyses are used for many critical purposes, including for statistical postprocessing and situational awareness (i.e., monitoring of current conditions). Such analyses commonly leverage a short-term numerical prediction to provide a background (first guess) forecast. In situations with a paucity of surface observations, the hourly surface temperature analysis may strongly reflect bias and

Corresponding author: Dr. Thomas M. Hamill, tom.hamill@noaa.gov

DOI: 10.1175/MWR-D-19-0028.1

error character of the background forecast. Hence, quantifying and benchmarking the errors in short-range surface temperature background forecasts used in rapidly updating data assimilation procedures is a useful first step in evaluating options for how to make progress in improving hourly surface-temperature analyses.

Following the benchmarking work of others, notably Best et al.'s (2015) benchmarking of surface latent and sensible heat flux, we have developed a benchmark for 1-hourly surface-temperature forecasts. In Hamill 2020, hereafter Part I), the reference benchmark was a station-based model of 1-h surface temperature forecasts. For surface weather stations in the contiguous United States (CONUS) with relatively long and complete records, an hourly and seasonally dependent climatology of surface temperature was developed for each station. The current hour's deviation from that climatology was persisted for 1 h and added to the climatology for the next hour to generate the synthetic 1-h benchmark forecast. Part I showed that this station benchmark had substantially reduced errors and bias when compared with raw HRRR forecasts. July 2015 root-mean-square errors (RMSEs) were approximately half in the station benchmark relative to 1-h forecasts of surface temperature from the raw High-Resolution Rapid Refresh System (HRRR; Benjamin et al. 2016). The August 2018 raw HRRR RMSEs and biases were substantially improved relative to the station benchmark but were still larger.

Surface temperatures are extremely spatially heterogeneous. Due to subgrid-scale differences in elevation, terrain type, shading, precipitation, soil and vegetation type, snow cover, and more, surface temperatures can vary substantially even on scales of hundreds of meters. Hence, a legitimate critique of Part I is that the low error of its benchmark may be due to its ability to represent persistent subgrid variability in the verification data that numerical model guidance cannot represent without much finer horizontal resolution.

This second part in the series describes and applies an alternative benchmark, one based on a gridded statistical interpolation of the observations' deviation from a seasonally and diurnally dependent climatology. This procedure will be called the ''gridded benchmark'' hereafter. To make sure that the potential advantage the station benchmark had in representing site-specific variability is avoided, a cross-validation procedure is used. When producing the gridded benchmark forecasts at a particular station, that station's data are excluded; the cross-validation procedure is explained more in section 2d. The operating hypothesis of this article is that even this more rigorous benchmark will still provide competitive errors relative to numerical weather predictions in regions with moderate to dense station observations. That is, the advantages from leveraging the current analyzed deviation from climatology are hypothesized to equal or outweigh the benefits of being able to predict dynamical changes due to meso and synoptic-scale weather variations with a NWP system, which may have short-term systematic error tendencies larger than the weather-induced tendencies.

The procedure for generating the gridded benchmark forecast is relatively simple and is illustrated in Fig. 1. The procedure requires surface-temperature climatology grids that are seasonally and hourly dependent. It also requires a procedure for statistically interpolating the current hour's surface-temperature observations to create a gridded estimate of the anomalies from climatology. The benchmark is generated by statistically interpolating the current hour's stations' anomalies from climatology to create a gridded field of anomalies. This anomaly field is persisted for 1 h (i.e., a 1-h forecast is generated by adding the persisted anomaly to the next hour's climatology appropriate to that day of the year).

The gridded seasonally and hourly varying climatologies are based upon high-resolution climatologies of daily maximum and minimum surface temperature over the CONUS produced by the Parameter-elevation Relationships on Independent Slopes Model (PRISM) group at Oregon State University (Daly et al. 2008 and references therein). PRISM monthly daily maximum and minimum climatologies are synthesized with the seasonally and diurnally dependent station climatologies (Part I) to generate the high-resolution, gridded, seasonally and diurnally dependent climatologies.

The statistical interpolation of anomalies from the climatology are produced via optimal (statistical) interpolation (Gandin 1965 and Daley 1991, chapter 4). While there are several other methods in the literature for creating observation-based surface temperature analyses such as Uboldi et al. (2008), Haylock et al. (2008), and Lussana et al. (2018), the procedure used here is simple and easy to apply.

Section 2 below will describe the datasets and the methods that will be used in evaluating the numerical guidance with respect to the gridded benchmark. Section 3 will describe: 1) the numerical procedures used for generating the gridded benchmark, including the procedure for developing a gridded climatology unique to each hour of the day and Julian day of the year; 2) the optimal interpolation procedure for producing the gridded analysis of the deviations from climatology, and 3) their combination to generate the statistical 1-h background forecast. Section 4 will discuss the results. Section 5 provides a discussion on the broader applicability of the

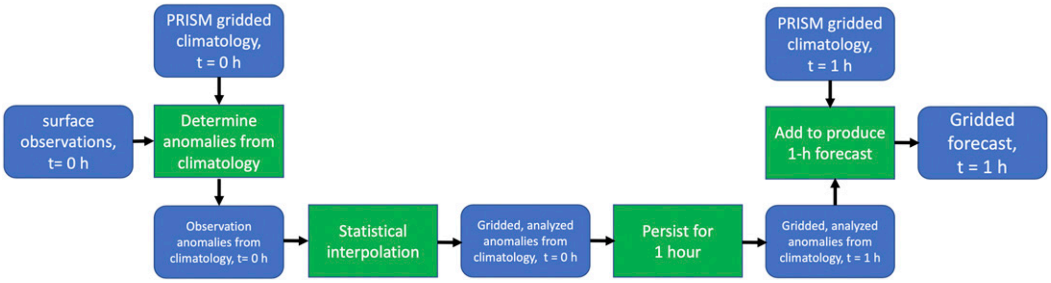
PRISM gridded climatology, t = 0 h
surface observations, t= 0 h
Determine anomalies from climatology
Observation anomalies from climatology, t= 0 h
Statistical interpolation
Gridded, analyzed anomalies from climatology, t= 0 h
Persist for 1 hour
Gridded, analyzed anomalies from climatology, t = 1 h
Add to produce 1-h forecast
PRISM gridded climatology, t = 1 h
Gridded forecast, t = 1 h

FIG. 1. Illustration of the procedure for generating 1-h benchmark gridded forecasts of surface temperature. Blue boxes are datasets. Green boxes are algorithmic procedures.

methodologies developed here. Section 6 provides conclusions.

2. Data and evaluation methods

The gridded benchmarking procedure will generate statistical 1-h forecasts over the CONUS based on two data sources, high-resolution monthly gridded surface-temperature climatologies and time series of CONUS station observations. Their synthesis will provide the gridded benchmark for dynamical 1-h forecasts from the raw HRRR system in July 2015 and August 2018, absent the elevation corrections for analysis grid differences. These data are now described along with forecast evaluation methods.

a. PRISM surface temperature climate data

Recognizing great user demand for high-resolution gridded climatologies of key variables such as surface temperature, humidity, and precipitation, the PRISM group at Oregon State University developed high-resolution gridded climatologies for domains such as the CONUS (Daly et al. 2008). (They provide selected datasets to the community free of charge, available for download through their web portal, http://prism.oregonstate.edu/normals/.) We downloaded high-resolution (0.04667°) gridded climatologies of maximum and minimum surface temperatures for each month of the year over the CONUS. A cubic-spline temporal fitting was applied to each grid point to develop gridded daily maximum and minimum climatologies from the monthly climatologies; this procedure will be described in section 3.

b. Surface-temperature observations

As in Part I, the observation dataset used in this experiment for developing the benchmark and validating the forecasts was the National Center for Atmospheric Research (NCAR) dataset 472.0, an archive of quality-controlled hourly surface observations over North America (Meteorological Development Laboratory/Office of Science and Technology/National Weather Service/NOAA/U.S. Department of Commerce 1987). Surface temperatures over the CONUS were used for the period 0000 UTC 1 January 2004–2300 UTC 28 January 2019. We chose to further limit use of surface temperatures in this dataset to only those with observation sites where data were available at 97% or more of the hours, days, and years in the analysis period. This observation availability cutoff was made based on the importance of an accurate estimation of the climatology to this procedure. 1118 stations were available in the CONUS.

c. Experimental HRRR 1-h forecasts

To compare the gridded benchmark against numerical forecasts, 1-h forecasts of raw background surface temperatures were extracted from the HRRR limited-area prediction system of Benjamin et al. (2016), again absent the elevation adjustment for RTMA grid differences. For comparison with station data, the HRRR forecast value at the nearest grid point was used. HRRR version 1 forecast data were used in July 2015 and HRRR version 3 data were used in August 2018. The HRRR system generated hourly analyses and numerical forecast guidance to +15 h lead time. HRRR data are used for many applications in the NWS, including severe weather prediction, short-term precipitation prediction, aviation applications, and for providing background fields in the generation of an ‘‘analysis of record’’ (De Pondeca et al. 2011). The underlying prediction system was the Advanced Research version of the Weather Research and Forecasting (WRF) Model (WRF-ARW), with a 3D ensemble–variational data assimilation system. See Benjamin et al. (2016) for more details.

HRRR forecasts in July 2015 were sometimes unavailable; in particular, 1-h forecasts initialized for the following dates: 1500 UTC 1 July 2015, 1000 UTC 2 July 2015, 0500 UTC 3 July 2015, 1400 UTC 3 July 2015, 1400 UTC 5 July 2015, 1200 UTC 6 July 2015, 2100 UTC

6 July 2015, 0800 8 July 2015, 0300 UTC 10 July 2015, 0800 UTC 11 July 2015, 1600 UTC 11 July 2015, 1400 UTC 18 July 2015, 2100 UTC 18 July 2015, 1300 UTC 22 July 2015, 1100 UTC 23 July 2015, 1300 UTC 26 July 2015, and 2100 UTC 28 July 2015. The validation of both the HRRR and the gridded benchmark will not include data at these times.

d. Evaluation methods

Standard methods of evaluation of deterministic forecasts were used, including root-mean-square error (RMSE), mean absolute error (MAE), and bias, all following standard definitions in Wilks (2011). All grid points with observations within the CONUS mask were used in the evaluation. The 5th and 95th percentile confidence intervals of a distribution consistent with the null hypothesis of no differences are provided on the comparative plot of errors from the two systems. The confidence intervals were determined through a paired block bootstrap algorithm following Hamill (1999). Statistics were calculated separately for each day, and statistics from one day to the next were assumed to be independent (ibid).

The gridded benchmark 1-h forecasts were cross validated; HRRR data were not. Specifically, 10 replications were produced. In each replication 90% of the observations were used to generate the gridded benchmark, and for the remaining 10% of the sites the observations 1 h later were used for validation. Over the 10 replications, verification against the full set of quality controlled CONUS observations were thus performed. The cross-validated results should provide a realistic benchmark for the HRRR forecasts; both suffer the same issues of not representing the site-specific variability of the verifying observations. Since the observation network was aggressively thinned in this study to include only those sites with nearly complete time series, we anticipate that many more observations would be available to the gridded benchmark in an operational environment, improving its quality. The number of observations used in producing the PRISM climatology, for instance, is roughly an order of magnitude larger.

3. Numerical methods used to create the gridded benchmark

The 1-h forecast gridded benchmarks were created by generating a gridded statistical interpolation of the current hour's observed anomalies from the climatology for that hour and that Julian day. These deviations were persisted for 1 h and added to the next hour's climatology to generate the benchmark forecasts. The overall procedure is illustrated in Fig. 1.

To produce the benchmark forecast, we require seasonally and hourly dependent, high-resolution gridded climatologies and a procedure for generating a statistical interpolation of the current hour's deviations from that climatology. These are now described.

a. Development of hourly gridded CONUS surface temperature climatologies

The first step in creating the hourly and seasonally dependent surface-temperature climatology was to estimate gridded maximum and minimum temperature climatologies unique to each Julian day. This was achieved by cubic spline-fitting the gridded PRISM climatologies of monthly estimated of the daily maximum and minimum temperature to each Julian day of the year. Monthly PRISM values were assigned to the Julian day associated with the middle Julian day in each calendar month. These monthly values were also wrapped, repeating October–November–December data at the beginning of the time series and January–February–March data at the end of the time series. A cubic-spline function fitting (Press et al. 1992, their section 3.3) with 8 knots spaced throughout the calendar year was then used to estimate the maximum and minimum temperatures for each Julian day. This spline-fitting procedure was applied independently for each grid point in the CONUS. For each Julian day d, gridded arrays of PRISM-based maximum temperature $\mathbf{T}^{P}_{\max}(d)$ and minimum temperature $\mathbf{T}^{P}_{\min}(d)$ were thus created.

For the gridded benchmarking procedure, we required the gridded climatology at each hour of each Julian day, not only estimates of the daily maximum and minimum temperature climatology. The hourly station climatologies developed and discussed in Part I utilized were next utilized to estimate this. For each surface station s and for each hour h of day d, the hourly and seasonally dependent climatology $T_{\mathrm{clim}}(s, d, h)$ were then used to estimate that day's and hour's fractional value $F(s, d, h)$ between the daily minimum and daily maximum for each UTC hour $h \in [0, 23]$. For a chosen s and d, let $T_{\min}(s, d)$ represent the station's minimum climatological temperature over the full 24 h. Let $T_{\max}(s, d)$ be the station's maximum climatological temperature for that Julian day, and let $T_{\mathrm{clim}}(s, d, h)$ represent the climatological temperature at hour h, calculated in Part I. $F(s, d, h)$ was then defined as

$$F(s,d,h) = \frac{T(s,d,h) - T_{\min}(s,d)}{T_{\max}(s,d) - T_{\min}(s,d)}. \qquad (1)$$

The station fractional values $F(s, d, h)$ were objectively analyzed with a simple procedure (Cressman 1959) to create a gridded array estimate $\mathbf{F}(d, h)$ on the

0.04667° PRISM grid using a successive-corrections procedure with three passes and influence radii of 70, 50, and 30 grid points. For each Julian day and hour, the PRISM-based gridded climatology $\mathbf{T}^P(d,h)$ was then created as follows:

$$\mathbf{T}^P(d,h) = \mathbf{T}^P_{\min}(d) + \mathbf{F}(d,h) \times [\mathbf{T}^P_{\max}(d) - \mathbf{T}^P_{\min}(d)]. \qquad (2)$$

b. Generation of gridded analyses of the deviations from the climatology

The next step was to apply a data-assimilation (statistical interpolation) procedure to create a gridded analysis of the deviations from climatology $\mathbf{T}'_a(d, h)$ from differences between the observed surface temperatures and the climatology $\mathbf{T}^P(d, h)$. For this, optimal interpolation (Gandin 1965; Daley 1991, chapter 4) was used. The equation that was used to produce the analysis of the deviations from climatology was

$$\mathbf{T}'_a(d,h) = \mathbf{T}'_b(d,h) + \mathbf{B}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}\{\mathbf{t}_{\mathrm{stn}}(d,h) - \mathbf{H}[\mathbf{T}^P(d,h)]\}, \qquad (3)$$

where $\mathbf{T}'_b(d, h)$ was the deviation from background state, here an array of zeros (i.e., the background state was implicitly the seasonally, diurnally dependent PRISM-based climatology); $\mathbf{B}$ was the background error covariance matrix, describing the statistical relationships between deviations of the station observations from background climatology as a function of horizontal distance, vertical distance, and a coastal proximity index difference, described later; $\mathbf{H}$ was the forward operator, which for this high-resolution grid consisted of extracting the model forecast at the PRISM grid point closest to each observation; and $\mathbf{t}_{\mathrm{stn}}(d, h)$ was the observation column vector of station surface-temperature observations. The $\mathbf{H}[\mathbf{T}(d, h)]$ operator thus created a column vector of extracted climatological values of the same dimension as the observation vector at the station locations; $\mathbf{R}$ was the observation- and representativeness-error covariance matrix, with assumed error variance of $1\,C^2$, the same value used in hourly data assimilation procedure described in De Pondeca et al. (2011). Off-diagonal elements in $\mathbf{R}$ were zero, consistent with the assumption that observation errors at different stations are independent.

Analyzed deviations were then persisted for 1 h to create the forecast deviations. Let $\mathbf{P}(\cdot)$ = the persistence of the forecast anomaly model operator (i.e., an identity matrix). That is,

$$\mathbf{T}'_a(d,h+1) = \mathbf{P}[\mathbf{T}'_a(d,h)] = \mathbf{T}'_a(d,h). \qquad (4)$$

The final, full-field forecast is the PRISM climatology 1 h later added to the persisted anomaly to create the 1-h gridded benchmark forecast:

$$\mathbf{T}_a(d,h+1) = \mathbf{P}[\mathbf{T}'_a(d,h)] + \mathbf{T}^P(d,h+1). \qquad (5)$$

Let us return to the details of the optimal interpolation procedure. Spatial horizontal error-covariance models for upper-air data have often assumed that errors are horizontally homogeneous, isotropic, and Gaussian distributed with an estimated correlation length scale (e.g., Daley 1991, section 4.3). For surface data, after examining several alternatives, we decided to use an exponentially distributed spatial error correlation model (Rasmussen and Williams 2006) within $\mathbf{B}$, which better fit the summertime spatial relationships in the surface data, with its greater small-scale variability due to smaller-scale variations in land surface characteristics. During other seasons, the gamma-exponential covariance model often provided better fits (not shown). For two locations z_1 and z_2 with a horizontal great-circle distance d_h between them and with background-error variances $\sigma^2_{z_1}$ and $\sigma^2_{z_2}$ at the two locations, the exponential horizontal exponential covariance model is the product of the background error standard deviations at the two locations with the exponential correlation function:

$$\mathrm{Cov}(z_1, z_2) = \sigma_{z_1}\sigma_{z_2}\exp\left(-\frac{d_h}{\rho_h}\right). \qquad (6)$$

Here σ_{z_1} and σ_{z_2} denote the gridded estimates of background error spread (standard deviation), derived from an analysis of observation deviations from the hourly climatology, estimated from all stations in the CONUS domain for a given month. ρ_h is the objectively fitted horizontal length scale. Figure 2 provides a scatterplot of the horizontal error correlations between station pairs and the fitted correlation model for 0000 UTC. The observed Pearson correlations (dots) were computed from time series of vector pairs of the deviations of 0000 UTC observations from the 0000 UTC climatology. July and August data from 2004 to 2018 were used to generate this plot, and data were limited to location pairs east of 105°W longitude (i.e., east of the Rocky Mountains). Also, correlations were only computed if station elevation differences were less than 100 m. In this way, possible effects of terrain elevation and coastal proximity differences between station locations were more muted; incorporated of these effects will be discussed later. The correlation length scales were computed with the python software library function "scipy.optimize.curve_fit" and using the "Trust Region Reflective" minimization algorithm. As

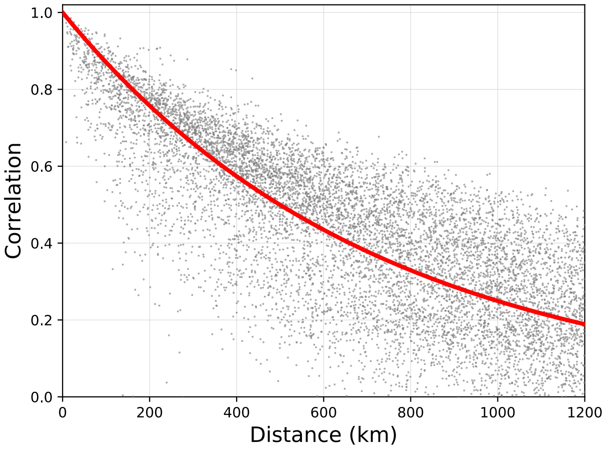

FIG. 2. Illustration of fitted exponential horizontal correlation function and a scatterplot of the correlation of pairs of vectors comprised of time series of the station observations' deviation from climatology. 0000 UTC July 2004–18 data were used to determine correlations between station pairs.

Fig. 2 shows, the exponential model provided a reasonable fit to station-to-station correlations. Though not shown here, station pairs with correlations less than the fitted line in Fig. 2 were more commonly found along the U.S. Gulf Coast and in the southeastern United States, where climatological background-error variances were also lower. Also not shown, we developed and tested a more complicated error covariance model that permitted the geographically varying horizontal length scales, leveraging these spatial variations in background error spreads as a predictor of the length scale. However, since the use of the resulting more complicated model did not reduce the analysis errors appreciably, this spatially varying horizontal length scale was omitted in the final algorithm for simplicity, and details are not described here.

Incorporation of two other factors into the spatial error covariance model slightly improved the estimation of covariances across the CONUS and the subsequent 1-h forecast error. First, as the algorithm must perform acceptably in mountainous regions as well as flatter regions, the distance norm in the error covariance model incorporated absolute differences in elevation. In this way, observations near mountain peaks had a greater impact on the effects of temperature analyses at other high elevations than observations in mountain valleys at a similar horizontal distance. The second factor we chose to incorporate was an index of the difference in "coastal proximity" between two locations. This was inspired by the use of a coastal proximity index produced in the PRISM project, whose use is explained in Daly et al. (2008). A map of the coastal proximity index used in this manuscript is shown in Fig. 3. Only values at grid points in the CONUS were used. Along the U.S. West Coast, temperature covariability is commonly related to differences in coastal proximity. Consider three observation sites, two along the coast and another inland, all equidistant. Covariances between the two coastal sites are commonly larger than covariances between either coastal site or the inland site. This is related to the limited inland penetration of marine air.

The three effects, horizontal distance, vertical distance, and differences in coastal proximity, were synthesized into a single distance norm. Let d_h represent the horizontal great-circle distance between the grid points in kilometers, normalized by its objectively fitted

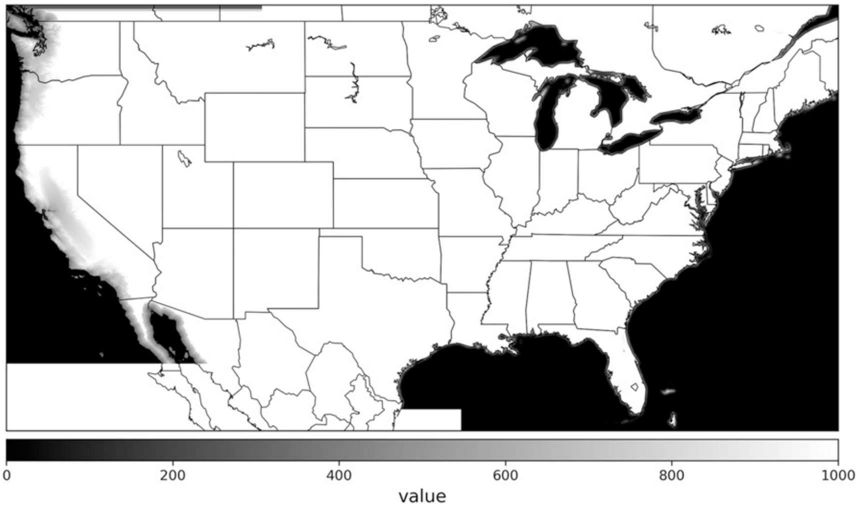

FIG. 3. Map of the coastal proximity index around the CONUS. Data courtesy of the PRISM group, Oregon State University. The index is used only at grid points within the CONUS.

horizontal length scale. Let d_v represent the vertical between the grid points in meters. This distance was normalized by an objectively fitted vertical correlation length scale ρ_v. The coastal proximity difference between two locations, $d_{\rm cp}$, was similarly normalized by an objectively fitted coastal-proximity index scale $\rho_{\rm cp}$. The distance norm between two locations was thus defined as

$$\|z_1 - z_2\| = \left[\left(\frac{d_h}{\rho_h}\right)^2 + \left(\frac{d_v}{\rho_v}\right)^2 + \left(\frac{d_{\rm cp}}{\rho_{\rm cp}}\right)^2\right]^{1/2}. \qquad (7)$$

The final background error covariance model between two locations was thus

$$\mathrm{Cov}(z_1, z_2) = \sigma_{z_1}\sigma_{z_2}\exp(-\|z_1 - z_2\|). \qquad (8)$$

The diurnal variations of the July fitted ρ_h, ρ_v, and $\rho_{\rm cp}$ length scales are shown in Fig. 4 using data from across the CONUS; only station pairs less than 2000 km distant from each other were used in the calculation. Vertical length scales were estimated to be shorter overnight, indicating a given vertical elevation difference between two locations will result in higher correlation by day and lower correlation by night. Perhaps this was related to the common pooling of cool air in mountain valleys overnight with concomitant decorrelation of valley temperatures with temperatures near adjacent mountain tops. Coastal proximity length scales peaked in the early morning hours for locations of relevance along the U.S. West Coast. Perhaps with overnight cooling, summertime U.S. West Coast land–sea temperature contrasts were at a minimum in the late-night hours, leading to the decreased influence of coastal proximity differences.

A gridded estimate of the standard deviations (spread) of observations with respect to the climatology are presented in for 0000 and 1200 UTC July 2015 and August 2018 data in Fig. 5. Such fields define the values of σ_{z_1} and σ_{z_2} in Eq. (8). To generate these gridded estimates, at each station and for each hour of the day, the standard deviations of the observations with respect to the observation climatology was determined using all July and August data. These station-based standard deviations were then objectively analyzed using a Cressman (1959) successive-corrections procedure with three passes and influence radii of 700, 400, and 250 km. The background for the first pass was a constant field reflecting the mean of all the calculated standard deviations at observation sites. The use of the successive-corrections procedure was arbitrary and chosen for its simplicity, but the accuracy of the subsequent analysis depends only slightly on the accuracy of these gridded estimates. The east–west spread differences in Fig. 5 at 0000 UTC may partly reflect the differing sun elevation angles; it was still late afternoon in the western United States but early evening in the eastern United States. Generally, daytime spreads were larger than overnight spreads as a consequence of day-to-day variability in solar-radiation reaching the surface, and concomitant variations day-to-day variability of surface heating. The lesser spread in the eastern United States may also have been a reflection of climatologically cloudier and moister atmospheric and soil

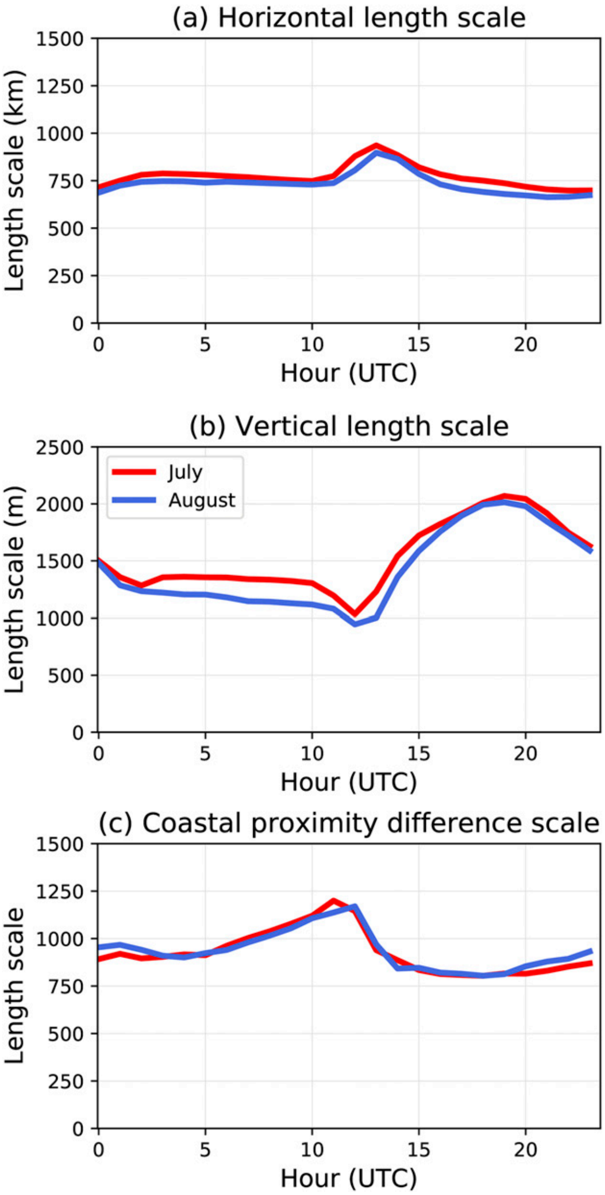

FIG. 4. Background error covariance fitted length scales for (a) horizontal distance, (b) vertical distance, and (c) coastal proximity difference. July values are red and August are blue.

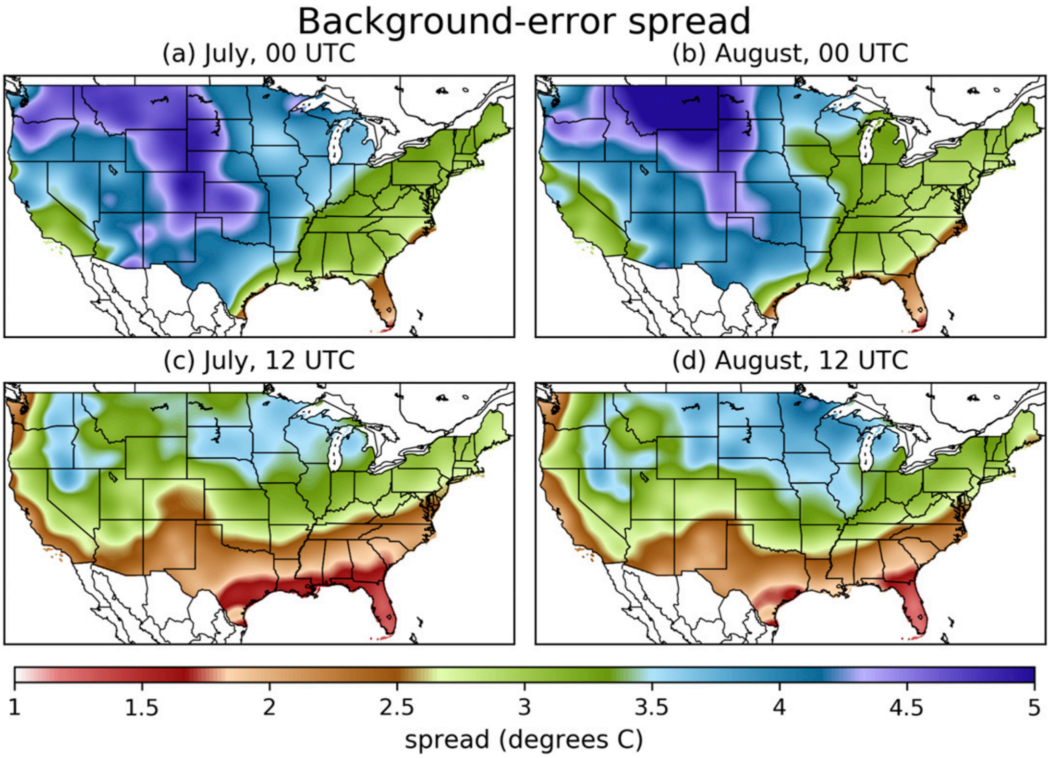

FIG. 5. Gridded estimates of background error spreads (the standard deviation of observed temperatures with respect to the climatology) across the CONUS, a component of the model of background error covariances. (a) July, 0000 UTC; (b) August, 0000 UTC; (c) July, 1200 UTC; and (d) August, 1200 UTC.

states and their constraints on temperature variation (Dai et al. 1999). Spreads were larger in the northern and central Rockies and western Great Plains, perhaps in part reflecting great temperature differences between afternoons with and without thunderstorms. Spread variations were more zonal in character at 1200 UTC and lower in the southern United States, reflecting the reduced variability in the subtropics relative to the extratropics.

We turn now from the components of the error covariance model to the numerical procedure for applying the statistical interpolation. This was generated per Eq. (3), with background error covariances calculated from Eqs. (7) and (8). The term $(\mathbf{HBH}^{\mathrm{T}} + \mathbf{R})^{-1}$ was first precalculated at a given time using all available observations. $\mathbf{HBH}^{\mathrm{T}}$ represented the climatological background error covariance matrix between all the observation locations. $\mathbf{HBH}^{\mathrm{T}} + \mathbf{R}$ was inverted through an $\mathbf{LU}$ decomposition procedure (Press et al. 1992), and $(\mathbf{HBH}^{\mathrm{T}} + \mathbf{R})^{-1}\{\mathbf{t}_{\mathrm{stn}}(d, h) - \mathbf{H}[\mathbf{T}(d, h)]\}$ was computed through $\mathbf{LU}$ backsubstitution (Press et al. 1992). The analysis procedure then looped over each grid point in the CONUS, calculating the cross covariance $\mathbf{BH}^{\mathrm{T}}$ between that grid point and every observation location, again with covariances calculated using Eqs. (7) and (8). The final analysis for that grid point was generated from the dot product of the associated row of $\mathbf{BH}^{\mathrm{T}}$ with $(\mathbf{HBH}^{\mathrm{T}} + \mathbf{R})^{-1}\{\mathbf{t}_{\mathrm{stn}}(d, h) - \mathbf{H}[\mathbf{T}(d, h)]\}$.

Two plots of the Kalman gain (Maybeck 1994, section 3.11), $\mathbf{BH}^{\mathrm{T}}(\mathbf{HBH}^{\mathrm{T}} + \mathbf{R})^{-1}$ are shown in Fig. 6, for 0000 UTC in July at Sacramento, California, and Denver, Colorado. These present the grid of multiplication factors to apply to the observation increment $\{\mathbf{t}_{\mathrm{stn}}(d, h) - \mathbf{H}[\mathbf{T}(d, h)]\}$ at the station location. Were that station the only one available, the analysis increment would simply be the Kalman gain times that single observation increment. At Sacramento, the observation increment is spread out more readily to other locations in the San Joaquin Valley than to points in the Sierra Nevada at a similar distance. For Denver, an observation has its greatest impact along the Front Range in Colorado and Wyoming, with progressively diminished impact to the west, in the high peaks of the Rocky Mountains.

4. Results

Figure 7 provides an example of the construction of the gridded benchmark from climatology and a persistence of the previous hour's analysis of anomalies from climatology. The time at which the analysis of

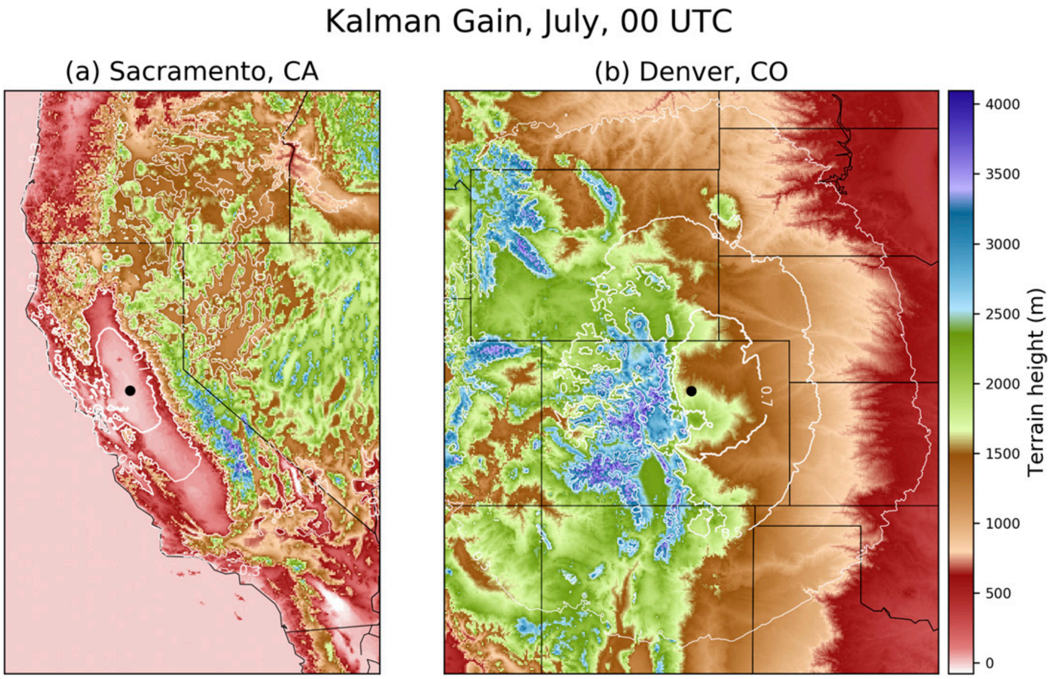

FIG. 6. Illustration of 0000 UTC July Kalman gain for (a) Sacramento, CA, and (b) Denver, CO. Stations are located at the black dots. Colors represent terrain elevation. Kalman gain contours are plotted for the 0.1, 0.3, 0.5, and 0.7 levels, with progressively thicker contours for each.

temperature deviations from climatology was generated was 0000 UTC 1 July 2015. From this, a 1-h persistence forecast of these anomalies is made for 0100 UTC 1 July 2015. The left-hand panel shows the PRISM-based climatology in the western United States for 0100 UTC 1 July 2015, the date and time of the 1-h forecast. It has significant spatial detail, including elevation-related temperature variations and cooler temperatures for grid points near the Pacific Ocean. The middle panel provides both the observed anomalies from the PRISM climatology at 0000 UTC (the overplotted numbers) and the resulting gridded analysis of the anomalies in colors. The analysis exhibited temperature variations that appear to fit the observed anomaly data well, and the gridded analysis had elevation- and coastal-proximity related spatial detail. Finally, the right-hand panel of Fig. 6 shows the final 1-h forecast, produced from the addition of the 0100 UTC climatology and the (persisted) 0000 UTC analyzed deviations.

A comparison of CONUS-averaged error statistics of the HRRR and the statistical benchmark are provided in Fig. 8. This may be compared this with Fig. 8 from Part I; now the benchmark is a cross-validated, gridded 1-h forecast as opposed to Part I's station-based benchmark. Like the station benchmark, the statistical benchmark appears to exhibit insignificant bias for all times of the day when averaged over many station locations. Errors were larger with the gridded benchmark than they were for the station benchmark in Part I.

The gridded benchmark errors were statistically significantly lower than those from 1-h HRRR forecasts for all hours of the day in July 2015. The differences were most striking during the daytime, ~1500–2300 UTC. As the sun rises, energy inputs become larger and if cloud cover is misforecast, then energy input errors to the surface radiation balance are also larger. In this situation the downward solar radiation reaching the surface will be misestimated, and the ground- and sensible heat flux will also likely be misestimated.

The August 2018 HRRR forecasts tell a different story; with many improvements to the system, the HRRR forecasts were very near and sometimes slightly lower in error than the gridded benchmark. August 2018 HRRR forecast bias was reduced substantially relative to the July 2015 values.

Are there specific geographic regions where the surface temperatures from the HRRR and the gridded benchmark were notably different from each other? The 0000 UTC HRRR and gridded benchmark RMS errors at stations across the CONUS are plotted in Fig. 9, with corresponding biases in Fig. 10. Figure 9a shows RMSE from the HRRR system of July 2015, and Fig. 9b the RMSE for the HRRR in August 2018.

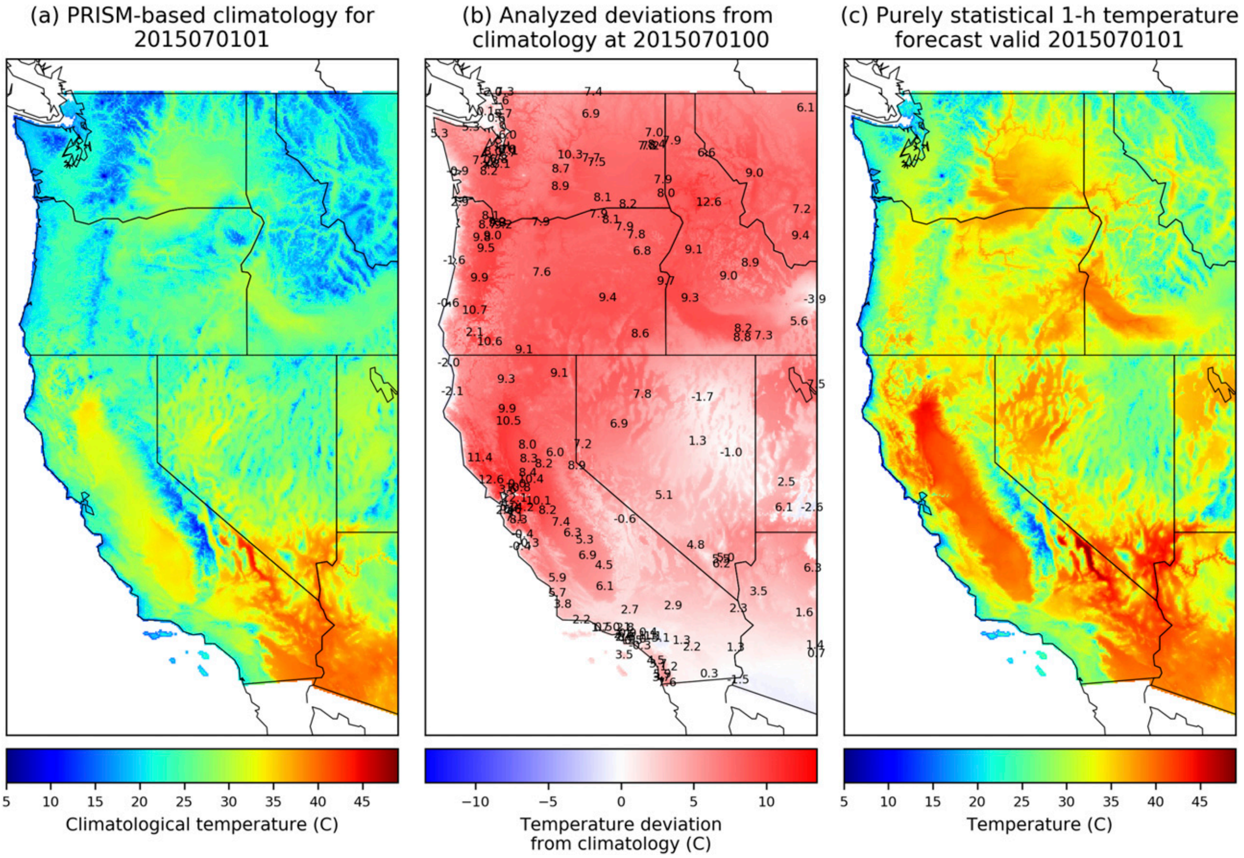
(a) PRISM-based climatology for 2015070101
(b) Analyzed deviations from climatology at 2015070100
(c) Purely statistical 1-h temperature forecast valid 2015070101
5 10 15 20 25 30 35 40 45
Climatological temperature (C)
−10 −5 0 5 10
Temperature deviation from climatology (C)
5 10 15 20 25 30 35 40 45
Temperature (C)

FIG. 7. An illustration of the procedure for generating gridded statistical 1-h forecasts of surface temperature, valid at 0100 UTC 1 Jul 2015. (a) The PRISM-based climatology adapted to the date of the forecast. (b) The analysis of temperature anomalies from the 0000 UTC climatology (colors) and the observed station anomalies (plotted numbers). (c) The sum of (a) and (b), a statistical forecast blending the climatology with the 1-h persisted anomaly.

Respective benchmark RMSEs are presented in Figs. 9c and 9d. For July 2015, errors were more often than not higher in the HRRR system, and these errors were particularly large in the mountainous western United States and northern Great Plains. The gridded benchmark, with a few exceptions, exhibited lower and more geographically uniform errors and smaller biases. However, by August 2018, HRRR errors were dramatically reduced, especially in the Rocky Mountains, where their errors were commonly lower than for the gridded benchmark. Biases for the HRRR system were again markedly reduced in August 2018 relative to the July 2015 values.

Figure 11 provides a scatterplot of the comparative RMSE and bias at 0000 UTC between the HRRR and the gridded benchmark. Again, for July 2015, the majority of verification sites have smaller RMSE in the gridded benchmark than the HRRR system. For August 2018, now the majority of sites have lower errors in the HRRR system relative to the gridded benchmark. August 2018 biases in the HRRR and the gridded benchmark were comparable in magnitude and often similar in sign. Returning to the original hypothesis, that this more rigorous gridded benchmark will still provide competitive errors relative to numerical weather predictions in regions with moderate to dense station observations, the hypothesis was confirmed for July 2015 data, but the benchmark was only somewhat competitive in August 2018.

5. Discussion on broader applicability

a. *Hourly high-resolution surface-temperature analyses and reanalyses*

This article has demonstrated that it is possible to generate accurate real-time and retrospective 1-h surface temperature forecasts in data-rich regions that have PRISM climatologies without the computational expense of a numerical weather prediction system. The gridded benchmark procedure generated forecasts that were more skillful than the raw HRRR system in July 2015 and generally competitive with the HRRR in

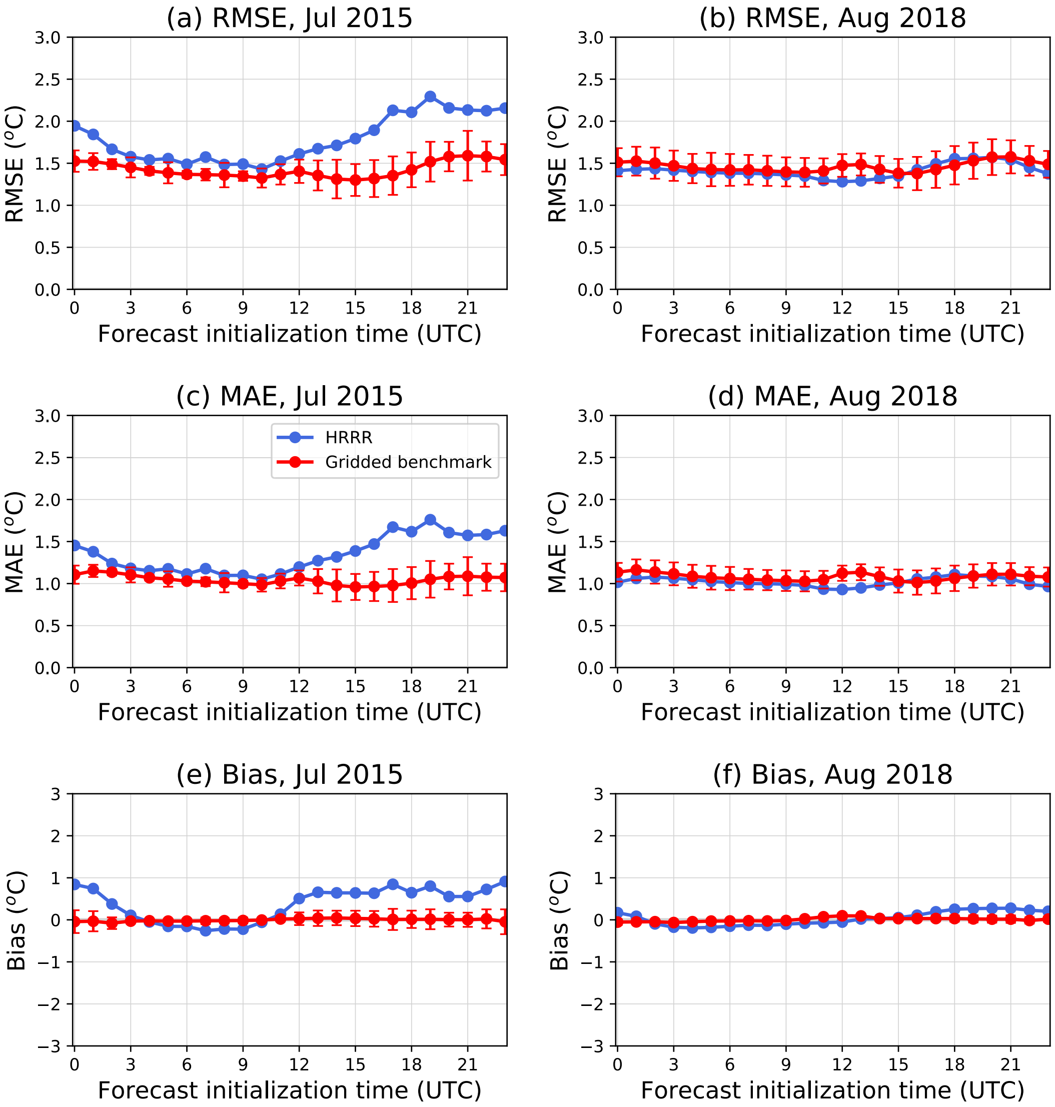

FIG. 8. 1-h surface temperature forecast error statistics for CONUS HRRR forecasts interpolated to stations and for the gridded benchmark at those stations. (a) Root-mean-square error, July 2015; (b) root-mean-square error, August 2018; (c) mean absolute error, July 2015; (d) mean absolute error, August 2018; (e) bias, July 2015, and (f) bias, August 2018. Error bars are recentered around the station benchmark and represent the 5th and 95th percentiles from a paired block bootstrap distribution consistent with the null hypothesis of no differences in mean.

August 2018. The HRRR background forecasts did not incorporate the RTMA procedure to adjust the background for differences in elevation between forecast and analysis grids.

This technology may facilitate the generation of computationally inexpensive multidecadal surface temperature reanalyses over the CONUS. Statistically generated 1-h forecasts could potentially be used as the replacement numerical model background states in the statistical interpolation analysis of the current hour's surface temperature observations. Similar statistical interpolation procedures to those used in the generation of the forecast benchmark could be used, but the analysis would likely use the 1-h gridded benchmark forecast for the background instead of climatology. The generation of a single analysis takes $O(1)$ min on a current-generation desktop computer. Hence, to produce 20 years of hourly background forecasts and 20 year of reanalyses would take roughly 2 × 20 years × 365 days × 24 h = 350 400 min, or 242 days on a desktop. With a 10-computer cluster, production would take less than one month. The procedure

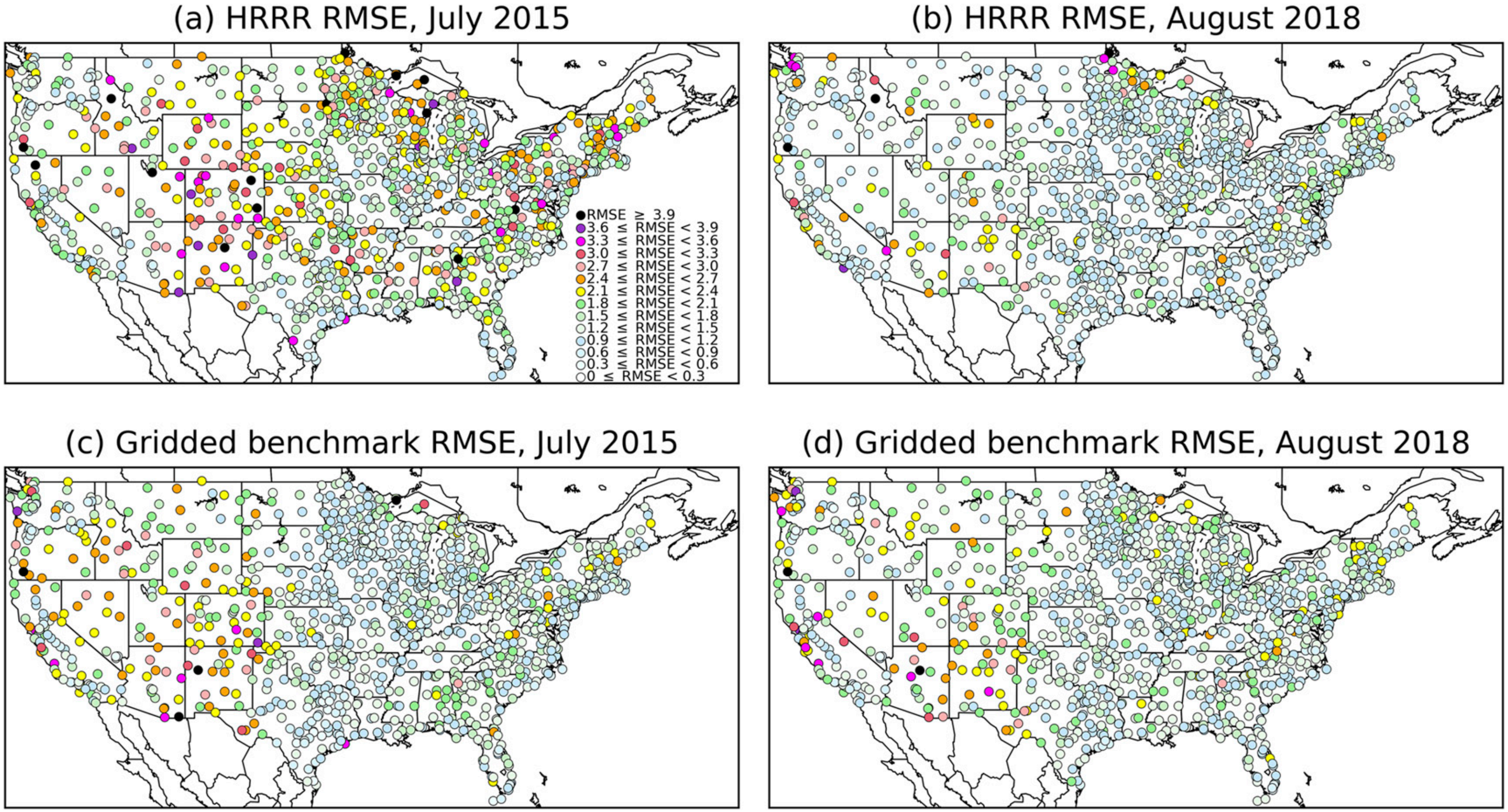

FIG. 9. RMSEs of 1-h July 2015 forecasts initialized at 0000 UTC. (a) July 2015 HRRR model forecasts; (b) August 2018 HRRR model forecasts; (c) July 2015 gridded benchmark; and (d) August 2018 gridded benchmark.

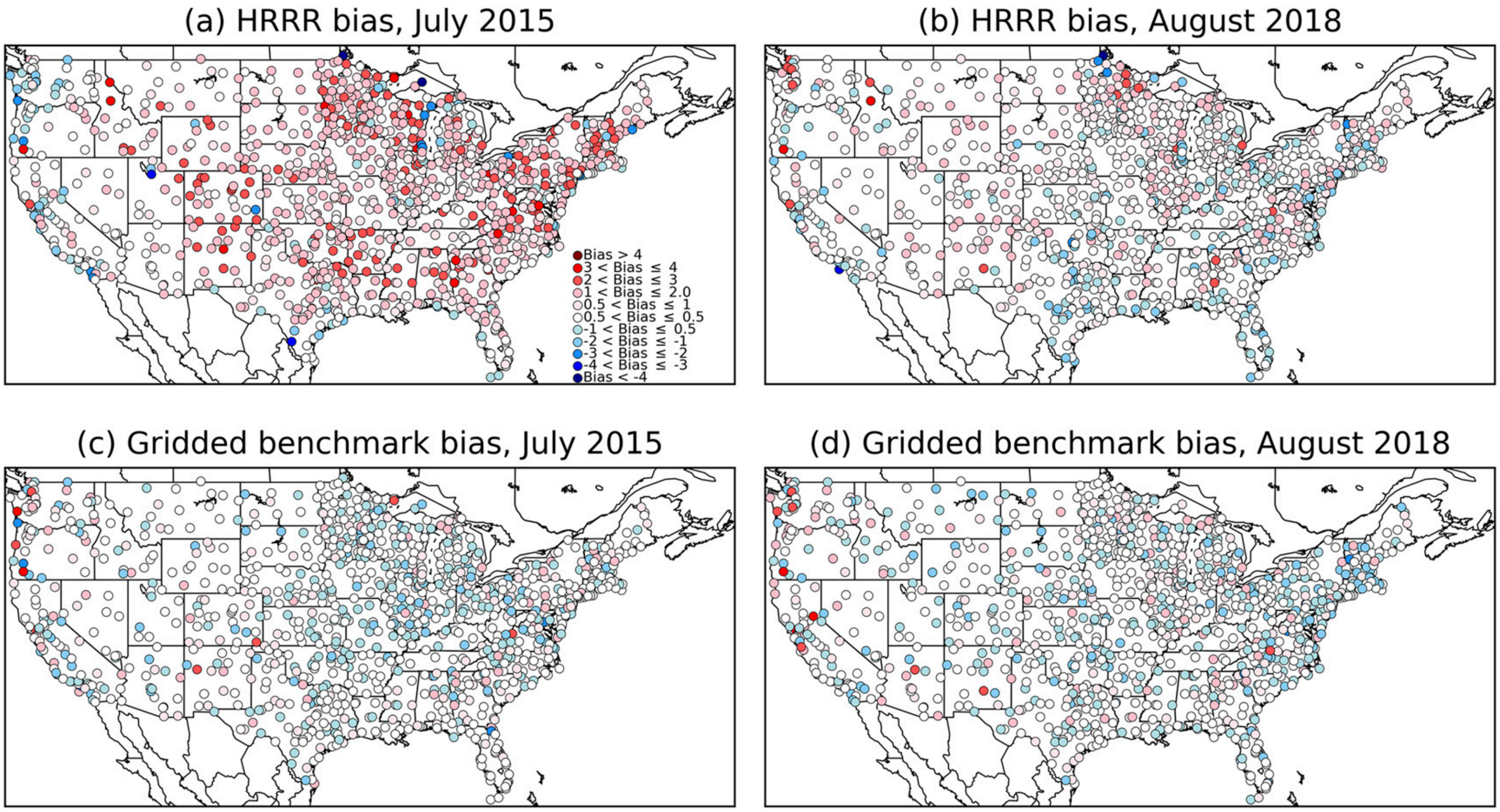

FIG. 10. Biases of 1-h July 2015 forecasts initialized at 0000 UTC. (a) July 2015 HRRR model forecasts; (b) August 2018 HRRR model forecasts; (c) July 2015 gridded benchmark; and (d) August 2018 gridded benchmark.

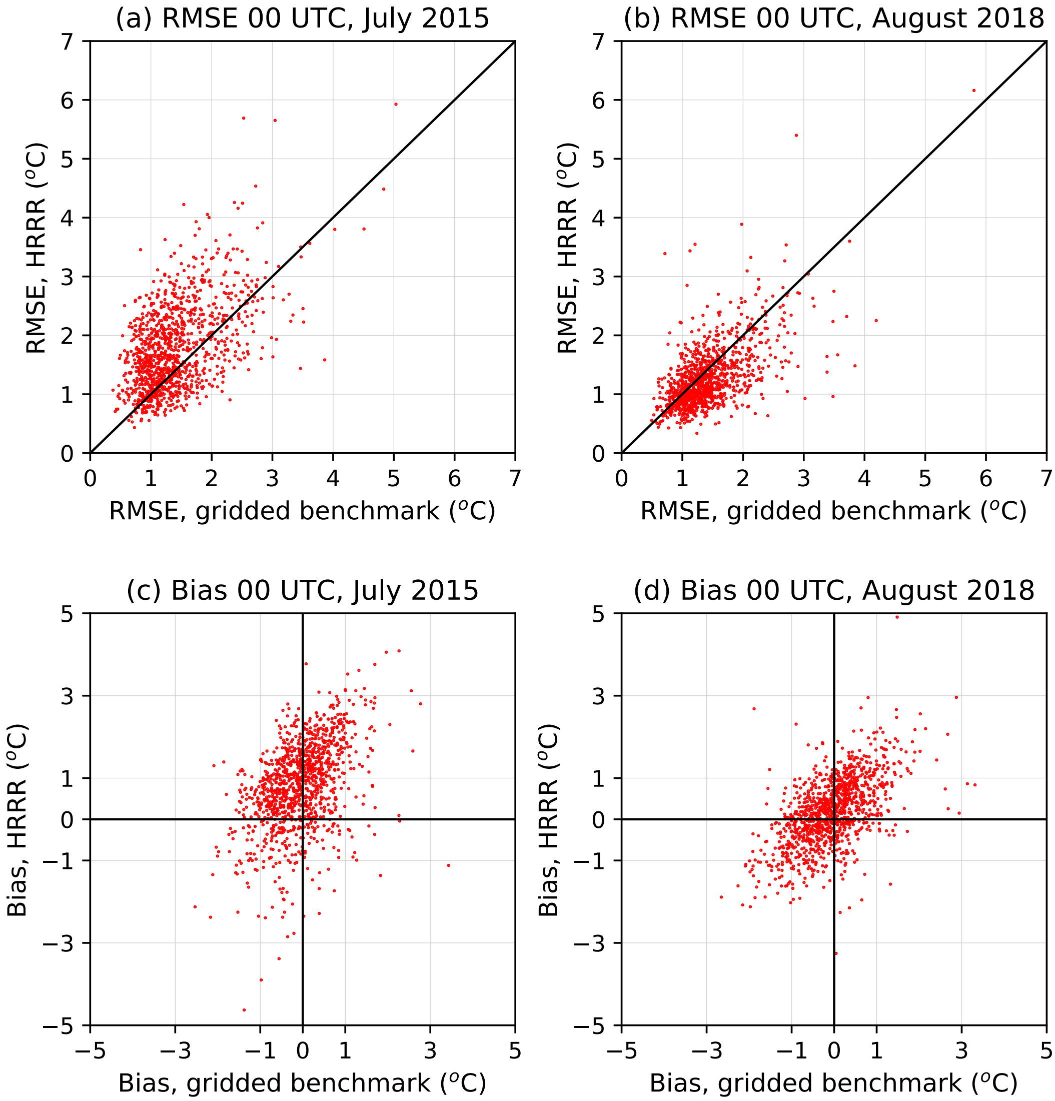

FIG. 11. Scatterplot of errors and bias for HRRR vs gridded benchmark. (a) 1-h forecast RMSE, July 2015; (b) 1-h forecast RMSE, August 2018; (c) 1-h forecast bias, July 2015, and (d) 1-h forecast bias, August 2018.

could also be applied in other regions where PRISM or PRISM-like climatologies are available and where observation density is adequate. The major tasks in performing such a reanalysis over the CONUS would likely include gathering and quality control of a more extensive set of surface-temperature observations and the reformulation of the background error covariance model for the actual analysis procedure.

The current gridded benchmark produced its forecasts on the 0.04667° PRISM grid, rather than on the 2.5-km grid used in the RTMA system that provides the ‘‘analysis of record’’ for the CONUS for the U.S. National Weather Service. There would be additional development needed to adapt the technology to this RTMA grid with its associated terrain elevation dataset.

b. Applications of the hourly surface-temperature analyses and reanalyses

Were accurate, unbiased reanalyses and real-time analyses available, several impactful applications could be anticipated. For example, the U.S. NWS uses gridded surface temperature analyses to statistically postprocess multimodel ensemble numerical forecast guidance to correct bias and generate improved deterministic and probabilistic forecast products. This is referred to as the National Blend of Models (NBM) project, and these postprocessed products are used to provide human forecasters with a gridded forecast estimate for use in product generation. Examples for precipitation forecast development can be found in Hamill et al. (2017) and Hamill and Scheuerer (2018). Surface temperature forecast products are one of the flagship elements in the NBM, and their accuracy depends in turn on the accuracy of the analyses against which they are calibrated. Improved surface temperature reanalyses over the CONUS and over a long period of time could be leveraged along with reforecasts (Hamill et al. 2013) to increase training sample size and improve the postprocessed NBM temperature guidance. It could also be used to improve longer-lead postprocessed guidance such as week +2 to week +4 temperature forecast products generated by the Climate Prediction Center. The forecast temperature training data in the current NBM use a decaying-average bias correction technique (Cui et al. 2012) that requires archival of only the recent most forecast and analysis. While this procedure is attractive from the standpoint of minimizing storage space, the algorithm has deficiencies which are discussed in Hamill (2018), and greater accuracy and realistic detail can be expected when longer training datasets are leveraged.

Other applications of surface temperature reanalyses could be envisioned as well. Whether it is appropriate or not, many climate change studies may use reanalysis data to inform inferences about the changes in climate. A statistical procedure such as this, demonstrably uncontaminated by background forecast model bias, could provide better quality data for such studies. Also, after major high-impact events such as heat waves or cold outbreaks, there is commonly a desire to understand the causes and to make quantitative statements about the relative effects of weather variability, boundary conditions (soil or ocean state), or climate change (e.g., Dole et al. 2011). Unbiased reanalyses would be helpful for such attribution studies, and the methodology used here should make high-resolution CONUS surface temperature reanalyses feasible.

c. Bias correction of numerical weather prediction hourly background forecasts

An underpinning assumption in most data assimilation procedures is that the background state is unbiased. As shown in Fig. 7, this assumption was commonly violated in July 2015, though bias was substantially improved with August 2018 forecasts. However, it may be possible to apply a bias correction to the background forecast, and a time series of unbiased hourly background forecasts and analyses could facilitate this. See Dee (2005) for several possible bias-correction procedures that could leverage such data.

d. Combining dynamical and statistical background state estimates

Another straightforward method for improving NWP 1-h background estimates would be to linearly combine the estimates from the NWP system together with those from the gridded benchmark described above; this may reduce the bias and error of the background forecast used in the data assimilation. Several potential issues would need to be addressed, such as how to handle a potential discontinuity on the U.S.–Canadian and U.S.–Mexican borders, given that PRISM climatologies are currently unavailable for Canada or Mexico.

e. Forcings for land surface state estimation

The Global Land Data Assimilation System (GLDAS; Rodell et al. 2004) and similar offline land-state estimations systems require surface temperature estimates as a model forcing. Forcing such systems with high-quality surface-temperature reanalyses may improve soil temperature state estimates in such systems. If they in turn are used in the initialization of the soil state for numerical weather predictions, this may lead to improved surface-temperature forecasts, since surface temperatures are strongly affected by the soil temperature.

6. Conclusions

This article continued Part I's exploration of a benchmarking procedure for hourly surface temperature forecasts created by rapidly cycling numerical weather prediction systems. In this article, an innovative statistical interpolation procedure was developed to combine information from a seasonally and diurnally dependent gridded climatology of surface temperature over the CONUS together with a gridded 1-h persisted analysis of anomalies from that climatology based on station observations. The procedure for generating the persisted analyses was the commonly used "optimal" or statistical interpolation. The novel aspect of the statistical interpolation procedure was the model for background error covariances of climatology. This used an exponential distance norm with an effective distance combining components for horizontal distance, absolute vertical distance, and absolute distance in a coastal proximity index. This novel error covariance model produced gridded anomaly estimates with significant spatial detail in mountainous regions and along the U.S. West Coast.

Results presented here showed that the statistically generated 1-h benchmark forecasts provided lower root-mean square errors, mean absolute errors, and bias over the CONUS for the July 2015 test period relative to the

HRRR system. For August 2018, the HRRR system was much improved, with greatly reduced bias and errors that often were slightly lower than those produced by the gridded benchmark. As expected, the errors of the gridded product in this Part 2 of the series were substantially higher than for the station-based benchmark in Part I. This was because the gridded benchmark was generated from independent data (i.e., stations other than the ones used for evaluation). In this way, any errors of representativeness particular to stations were eliminated.

There was a dramatic improvement in 1-h surface temperatures in the HRRR system from version 1 (July 2015) to version 3 (August 2018), described largely in Benjamin et al. (2016). The higher error and bias in the HRRR 1-h forecasts of 2015 highlights the challenge of making hourly numerical predictions of surface temperature. These forecasts are readily contaminated by the misestimation of soil surface temperatures, misestimation of cloud amount and cloud optical properties, and many other effects. Many changes were made to the HRRR system between versions, including the assimilation of cloud and precipitation hydrometeor information and the use of surface temperature and humidity innovations to make increments to the soil temperature and moisture; see also Mahfouf et al. (2009) and Lin and Pu (2018). From the results presented here, these changes had a dramatic and positive impact on the quality of short-term surface temperature forecasts.

Section 5 discussed the implications of this research beyond the benchmarking of 1-h forecasts. A minor variant of the procedure discussed here could be used to generate computationally inexpensive, unbiased reanalyses and real-time analyses of surface temperature over the CONUS (the real-time analyses would supplement, not replace the RTMA). This in turn may stimulate many applications, from facilitating improved statistical postprocessing of surface temperature to the bias correction of background forecasts in the NWP system. We hope this article stimulates broader interest and future collaborations to develop such applications.

Acknowledgments. Chris Daly of Oregon State University and the PRISM group at the Oregon State University are thanked for generating this highly useful dataset, for consultation on its proper use, and for sharing their coastal proximity index datasets. The **LU** decomposition and back-substitution routines were taken from the *Numerical Recipes* text (Press et al. 1992). Funding for this project was provided both by ESRL Physical Sciences Division base funding provided by NOAA's Office of Oceanic and Atmospheric Research and by a funding from the U.S. NWS Office of Science and Technology Integration through the Meteorological Development Lab, project number T8MWQML.P00. Eric James of the ESRL Global Systems Division is thanked for extraction of the HRRR data from the tape archive.

REFERENCES

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, https://doi.org/10.1175/MWR-D-15-0242.1.

Best, M. J., and Coauthors, 2015: The plumbing of land surface models: Benchmarking model performance. *J. Hydrometeor.*, **16**, 1425–1442, https://doi.org/10.1175/JHM-D-14-0158.1.

Cressman, G. P., 1959: An operational objective analysis system. *Mon. Wea. Rev.*, **87**, 367–374, https://doi.org/10.1175/1520-0493(1959)087<0367:AOOAS>2.0.CO;2.

Cui, B., Z. Toth, Y. Zhu, and D. Hou, 2012: Bias correction for global ensemble forecast. *Wea. Forecasting*, **27**, 396–410, https://doi.org/10.1175/WAF-D-11-00011.1.

Dai, A., K. E. Trenberth, and T. R. Karl, 1999: Effects of clouds, soil moisture, precipitation, and water vapor on diurnal temperature range. *J. Climate*, **12**, 2451–2473, https://doi.org/10.1175/1520-0442(1999)012<2451:EOCSMP>2.0.CO;2.

Daley, R., 1991: *Atmospheric Data Analysis*. Cambridge Press, 472 pp.

Daly, C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. A. Pasteris, 2008: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int. J. Climatol.*, **28**, 2031–2064, https://doi.org/10.1002/joc.1688.

De Pondeca, M., G. S. Manikin, G. DiMego, S. G. Benjamin, D. F. Parrish, and R. J. Purser, 2011: The real-time mesoscale analysis at NOAA's National Centers for Environmental Prediction: Current status and development. *Wea. Forecasting*, **26**, 593–612, https://doi.org/10.1175/WAF-D-10-05037.1.

de Rosnay, P., G. Balsamo, C. Albergel, J. Muñoz-Sabater, and L. Isaksen, 2014: Initialisation of land surface variables for numerical weather prediction. *Surv. Geophys.*, **35**, 607–621, https://doi.org/10.1007/s10712-012-9207-x.

Dee, D. P., 2005: Bias and data assimilation. *Quart. J. Roy. Meteor. Soc.*, **131**, 3323–3343, https://doi.org/10.1256/qj.05.137.

Dirmeyer, P. A., and Coauthors, 2018: Verification of land–atmosphere coupling in forecast models, reanalyses, and land surface models using flux site observations. *J. Hydrometeor.*, **19**, 375–392, https://doi.org/10.1175/JHM-D-17-0152.1.

Dole, R., and Coauthors, 2011: Was there a basis for anticipating the 2010 Russian heat wave? *Geophys. Res. Lett.*, **38**, L06702, https://doi.org/10.1029/2010GL046582.

Gandin, L., 1965: *Objective Analysis of Meteorological Fields*. Israel Program for Scientific Translation, 242 pp.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

——, 2018: Practical aspects of statistical postprocessing. *Statistical Postprocessing of Ensemble Forecasts,* S. Vannitsem, D. Wilks, and J. Messner, Eds., Elsevier Press, 187–217.

——, 2020: Benchmarking the raw model-generated background forecast in rapidly updated surface temperature analyses. Part I:

Stations. *Mon. Wea. Rev.*, **148**, 689–700, https://doi.org/10.1175/MWR-D-19-0027.1.

——, and M. Scheuerer, 2018: Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing. *Mon. Wea. Rev.*, **146**, 4079–4098, https://doi.org/10.1175/MWR-D-18-0147.1.

——, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

——, E. Engle, D. Myrick, M. Peroutka, C. Finan, and M. Scheuerer, 2017: The U.S. national blend of models for statistical postprocessing of probability of precipitation and deterministic precipitation amount. *Mon. Wea. Rev.*, **145**, 3441–3463, https://doi.org/10.1175/MWR-D-16-0331.1.

Haylock, M., N. Hofstra, A. Klein-Tank, E. Klok, P. Jones, and M. New, 2008: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950-2006. *J. Geophys. Res.*, **113**, D20119, https://doi.org/10.1029/2008JD010201.

Holmes, T. R. H., T. J. Jackson, R. H. Reichle, and J. Basara, 2012: An assessment of surface soil temperature products from numerical weather prediction models using ground-based measurements. *Water Resour. Res.*, **48**, W02531, https://doi.org/10.1029/2011WR010538.

Lin, L., and Z. Pu, 2018: Characteristics of background error covariance of soil moisture and atmospheric states in strongly coupled land–atmosphere data assimilation. *J. Appl. Meteor. Climatol.*, **57**, 2507–2529, https://doi.org/10.1175/JAMC-D-18-0050.1.

Lussana, C., O. E. Tvieto, and F. Uboldi, 2018: Three-dimensional spatial interpolation of 2 m temperature over Norway. *Quart. J. Roy. Meteor. Soc.*, **144**, 344–364, https://doi.org/10.1002/qj.3208.

Mahfouf, J.-F., K. Bergaoui, C. Draper, F. Bouyssel, F. Taillefer, and L. Taseva, 2009: A comparison of two offline soil analysis schemes for assimilation of screen level observations. *J. Geophys. Res.*, **114**, D08105, https://doi.org/10.1029/2008JD011077.

Maybeck, P. S., 1994: *Stochastic Models, Estimation, and Control.* Vol. 1, Navtech Book and Software, 423 pp.

Meteorological Development Laboratory/Office of Science and Technology/National Weather Service/NOAA/U.S. Department of Commerce, 1987: TDL U.S. and Canada surface hourly observations (updated half-yearly). Research Data Archive, National Center for Atmospheric Research, Computational and Information Systems Laboratory, accessed 15 April 2019, http://rda.ucar.edu/datasets/ds472.0/.

Paquin-Ricard, D., C. Jones, and P. A. Vaillancourt, 2010: Using ARM observations to evaluate cloud and clear-sky radiation processes as simulated by the Canadian Regional Climate model GEM. *Mon. Wea. Rev.*, **138**, 818–838, https://doi.org/10.1175/2009MWR2745.1.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in Fortran*. 2nd ed. Cambridge University Press, 963 pp.

Räisänen, P., and H. Järvinen, 2010: Impact on cloud and radiation scheme modifications on climate simulated by the ECHAM5 atmospheric GCM. *Quart. J. Roy. Meteor. Soc.*, **136**, 1733–1752, https://doi.org/10.1002/qj.674.

Rasmussen, C. E., and C. K. I. Williams, 2006: *Gaussian Processes for Machine Learning*. MIT Press, 248 pp.

Rodell, M., and Coauthors, 2004: The Global Land Data Assimilation System. *Bull. Amer. Meteor. Soc.*, **85**, 381–394, https://doi.org/10.1175/BAMS-85-3-381.

Ruiz-Arias, J. A., C. Arbizu-Barrena, F. J. Santos-Alamillos, J. Tovar-Pescador, and D. Pozo-Vázquez, 2016: Assessing the surface solar radiation budget in the WRF model: A spatiotemporal analysis of the bias and its causes. *Mon. Wea. Rev.*, **144**, 703–711, https://doi.org/10.1175/MWR-D-15-0262.1.

Sandu, I., A. Beljaars, P. Bechtold, T. Mauritsen, and G. Balsamo, 2013: Why is it so difficult to represent stably stratified conditions in Numerical Weather Prediction (NWP) models? *J. Adv. Model. Earth Syst.*, **5**, 117–133, https://doi.org/10.1002/jame.20013.

Thelen, J.-C., and J. M. Edwards, 2013: Short-wave radiances: Comparison between SEVIRI and the Unified Model. *Quart. J. Roy. Meteor. Soc.*, **139**, 1665–1679, https://doi.org/10.1002/qj.2034.

Uboldi, F., C. Lussana, and M. Salvati, 2008: Three-dimensional spatial interpolation of surface meteorological observations from high-resolution local networks. *Meteor. Appl.*, **15**, 331–345, https://doi.org/10.1002/met.76.

Van Weverberg, K., C. J. Morcrette, H.-Y. Ma, S. A. Klein, and J. C. Petch, 2015: Using regime analysis to identify the contribution of clouds to surface temperature errors in weather and climate models. *Quart. J. Roy. Meteor. Soc.*, **141**, 3190–3206, https://doi.org/10.1002/qj.2603.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Yang, F., H.-L. Pan, S. K. Krueger, S. Moorthi, and S. J. Lord, 2006: Evaluation of the NCEP Global Forecast System at the ARM SGP Site. *Mon. Wea. Rev.*, **134**, 3668–3690, https://doi.org/10.1175/MWR3264.1.